A Multiple Imputation Score Test for Model Modification in Structural Equation Models

Maxwell Mansolf, Terrence D. Jorgensen, and Craig K. Enders

Author Note

Maxwell Mansolf, University of California – Los Angeles, Department of Psychology

Terrence D. Jorgensen, University of Amsterdam – Department of Child Education and Development

Craig K. Enders, University of California – Los Angeles, Department of Psychology

This work was supported by Institute of Educational Sciences award R305D150056

Published in Psychological Methods, 2020.

Correspondence concerning this article should be addressed to Maxwell Mansolf, Department of Psychology, 502 Portola Plaza, Los Angeles, CA 90095.

Contact: mamansolf@gmail.com

Abstract

Structural equation modeling (SEM) applications routinely employ a trilogy of significance tests that includes the likelihood ratio test, Wald test, and score test or modification index. Researchers use these tests to assess global model fit, evaluate whether individual estimates differ from zero, and identify potential sources of local misfit, respectively. This full cadre of significance testing options is not yet available for multiply imputed data sets, as methodologists have yet to develop a general score test for this context. Thus, the goal of this paper is to outline a new score test for multiply imputed data. Consistent with its complete-data counterpart, this imputation-based score test provides an estimate of the familiar expected parameter change statistic. The new procedure is available in the R package `semTools` and naturally suited for identifying local misfit in SEM applications (i.e., a model modification index). The article uses a simulation study to assess the performance (Type I error rate, power) of the proposed score test relative to the score test produced by full information maximum likelihood (FIML) estimation. Owing to the two-stage nature of multiple imputation, the score test exhibited slightly lower power than the corresponding FIML statistic in some situations but was generally well-calibrated.

Keywords: Multiple imputation, score test, modification index, expected parameter change, missing data

A Multiple Imputation Score Test for Model Modification in Structural Equation Models

Buse (1982) described a "trilogy of tests" for null hypothesis significance testing: the Wald test (Wald, 1943), the likelihood ratio test (Wilks, 1938), and the score test, also known as the Lagrange multiplier test or modification index (Rao, 1948; Saris, Satorra, & Sörbom, 1987; Sörbom, 1989). Our focus is the application of these tests in structural equation modeling (SEM) applications, particularly the score test as described by Bollen (1989, pp. 292–296). Despite their asymptotic equivalence as general hypothesis tests, researchers tend to use the tests for different purposes in SEM. The Wald test, for instance, is often used to evaluate the statistical significance of parameters in the fitted model, and univariate Wald *z* tests are routinely provided for each estimated parameter by SEM software as part of standard statistical output (e.g., Muthén & Muthén, 1998–2017; Rosseel, 2012). The likelihood ratio test is most commonly used to evaluate model fit or compare two nested models (e.g., the hypothesized model versus a saturated and/or baseline model; Bollen, 1989), and the test statistic is also used to construct comparative fit indices (Bentler & Bonett, 1980). Last, but certainly not least, the score test appears most often in structural equation modeling as the *modification index*, a statistic used to quantify the change in model fit that would result if a parameter constraint were freed during estimation (MacCallum, 1986; Sörbom, 1989). Although we use the terms "modification index" and "score test" interchangeably, the score test has many applications outside of structural equation modeling. For example, in econometrics (Godfrey, 1996) and genetics (Jaffrézic, White, & Thompson, 2003; Sato & Ueki, 2018), the score test is used both to identify inadequately specified models and to perform computationally efficient comparisons between two models.

Although researchers tend to apply the trilogy of tests in different ways and for different purposes, they are, in fact, exchangeable because they fundamentally compare two nested

models, albeit in different ways. The likelihood ratio test requires the researcher to explicitly fit two different models, whereas the Wald and score tests effectively use information from the hypothesized model to make projections about another model; the Wald test considers a more restrictive model where some of the estimated parameters are constrained to zero, whereas as the score test makes projections about a less restrictive model that adds paths or parameters. The differences in the useable patterns of the tests lie in their ease with which they can be applied to a particular task. For example, using the Wald statistic as a global test of model fit is hypothetically possible, but doing so would require exceedingly complex constraints on the saturated model parameters. More detailed discussions about the use of these statistics in an SEM context can be found in various SEM textbooks (e.g., Bollen, 1989, pp. 292–303; Kline, 2011, pp. 215–219; see also Chou & Huh, 2012).

In the context of incomplete data analyses, full information maximum likelihood (FIML) estimation (Allison, 1987; Arbuckle, 1996; Muthén, Kaplan, & Hollis, 1987) provides test statistics that are direct extensions of their complete-data counterparts. In fact, FIML significance tests have received considerable attention in the methodology literature and a great deal is known about these procedures (Kenward & Molenberghs, 1998; Savalei, 2010a, 2010b; Savalei & Bentler, 2009; Savalei & Yuan, 2009; Yuan, Tong, & Zhang, 2014; Yuan & Bentler, 2000, 2010; Yuan & Savalei, 2014; Yuan & Zhang, 2012). However, much less is known about test statistics for multiply imputed data. To date, much of the literature has focused on improving the small-sample properties of single degree of freedom Wald tests (Barnard & Rubin, 1999; Reiter, 2007; Steele, Wang, & Raftery, 2010), with relatively few studies investigating multiparameter versions of this test (Grund, Ludtke, & Robitzsch, 2016; Liu & Enders, 2016). We are aware of only two studies that investigate the application of imputation-based likelihood ratio tests to

SEM: Lee and Cai (2012) proposed a two-stage approach to computing this test that is analogous to two-stage FIML estimation (Savalei & Bentler, 2009; Savalei & Falk, 2014; Savalei & Rhemtulla, 2014; Yuan et al., 2014), and Enders and Mansolf (2018) examined the use of Meng and Rubin's (1992) pooled likelihood ratio statistic (also referred to as the D3 statistic in the literature; Schafer, 1997) as a global test of SEM fit. Importantly, methodologists have yet to develop a general score test for multiply imputed data, much less one that can serve as a so-called modification index for SEMs. As such, the goal of this paper is to outline a new score test procedure and use Monte Carlo computer simulations to evaluate its performance.

The score test and its use in model modification has been the source of considerable controversy in the SEM literature, much of which is warranted. Modification indices can be used to transform a poorly-fitting model into a well-fitting model by computing a score test with one degree of freedom for each fixed parameter (e.g., omitted path) in the model. For those parameters with large test statistics, the model can be iteratively re-estimated after lifting the relevant constraints until the desired level of fit is achieved. As many readers already know, the practice of data-driven model modification has been widely criticized in the literature (Bollen, 1989, pp. 300-303; Brown, 2014; Kaplan, 1990; MacCallum, Roznowski, & Necowitz, 1992; Yoon & Kim, 2014). While not diminishing these concerns, we note that it is the *usage* of modification indices that has been widely criticized, and not the test itself. Indeed, modification indices remain useful in *a priori*, *theoretically driven*, or *explicitly exploratory* model modification (MacCallum, 1986; see example applications in Byrne, Shavelson, & Muthén, 1989; Kwok, Luo, & West, 2010), and the use of modification indices for these purposes does not necessarily invalidate a model-building procedure. Further, the score test often performs similarly to the Wald and likelihood ratio tests when an SEM is approximately properly

specified, and the procedure has been recommended in cases where the asymptotic properties of other tests break down, such as testing the null hypothesis that a variance component is zero in a mixed effects model (Verbeke & Molenberghs, 2003). Because the score test only requires estimation of a restricted model, it can also be useful in situations where a less restrictive model is difficult to estimate or fails to converge. Lastly, a natural byproduct of the score test is an expected parameter change (EPC) statistic (Kaplan, 1989; Saris et al., 1987) that estimates the value of a parameter (or set of parameters) that would result from freeing model constraints (e.g., adding a path or set of paths to an SEM). Prior research (Saris, Satorra, & van der Veld, 2009; Whittaker, 2012) has demonstrated the promising performance of score tests—even in an exploratory fashion—when considered in combination with EPCs before modifying a hypothesized model. For these reasons, we argue that it is important to develop and evaluate a score test procedure for multiply imputed data.

The organization of this paper is as follows. First, we introduce notation and describe a pair of nested confirmatory factor analysis (CFA) models that we use to motivate and describe the proposed score test. Second, we provide a brief review of and rationale for multiple imputation in the SEM context. Third, we describe the score test for multiply imputed data. To ensure that the exposition is accessible to the broadest possible readership, we include online supplemental material that provides a concise description and summary of the maximum likelihood principles needed to understand the composition of the test statistic (Supplemental Appendix A). Fourth, we use Monte Carlo computer simulations to evaluate the imputation-based score test, comparing its performance to the FIML counterpart. Fifth, we include a real data analysis example that uses the R package `semTools` (Jorgensen, Pornprasertmanit,

Schoemann, & Rosseel, 2019) to apply the new test. Finally, we conclude with practical recommendations and avenues for future research.

Motivating Example and Notation

To place our ensuing description of the multiple imputation score test in a specific context, we rely on a simple CFA model with four continuous observed variables, $X_1, \ldots, X_4$. Figure 1a depicts the model as a path diagram. Figure 1b shows the same model with an added covariance between ε_1 and ε_2, the residuals associated with X_1 and X_2. For identification, the variance of the latent variable F is fixed to one and the mean of the latent variable is fixed to zero. This yields models with 12 (Figure 1a) and 13 (Figure 1b) parameters: four intercepts $\tau_1, \ldots, \tau_4$, which here estimate the means of $X_1, \ldots, X_4$; four factor loadings $\lambda_1, \ldots, \lambda_4$; four residual variances $\sigma^2_{\varepsilon_1}, \ldots, \sigma^2_{\varepsilon_4}$, a single residual covariance $\sigma_{\varepsilon_1\varepsilon_2}$ between ε_1 and ε_2 (Figure 1b only). Consider the case in which the fit of the model in Figure 1a is not satisfactory. To improve model fit, one may consider adding the residual covariance $\sigma_{\varepsilon_1\varepsilon_2}$ to the model. Using only the results from the model in Figure 1a, the score test can project the improvement in fit that would result from estimating the more complex model in Figure 1b. The test also provides the information required to estimate the value of the residual covariance that would result if this parameter were freed during estimation (i.e., the expected parameter change, or EPC).

More generally, consider the case in which a researcher compares two (parametrically) nested models[1], a *restricted model* and a *general model* (e.g., Figures 1a and 1b, respectively), with the intention of determining whether the general model provides a significantly better fit to

[1] By "parametrically nested", we mean that the restricted model is constructed by placing constraints on the parameters of the more general model. For brevity, we will simply use "nested" to refer to parametrically nested models.

the data than the restricted model. We let $\boldsymbol{\theta}_r = (\theta_1, \theta_2 \ldots, \theta_q)$ denote the vector of q parameters for the restricted model, and we let $\boldsymbol{\theta}_g = (\theta_1, \theta_2 \ldots, \theta_q, \theta_{q+1})$ denote the vector of $q + 1$ parameters of the general model, where the parameter θ_{q+1} in the general model is constrained to zero in the restricted model. We use the *g* and *r* subscripts to differentiate various quantities or features of these two models. Returning to the models in Figure 1, $\boldsymbol{\theta}_g$ for the general model (Figure 1b) is given by $\boldsymbol{\theta}_g = (\tau_1, \ldots, \tau_4, \lambda_1, \ldots, \lambda_4, \sigma^2_{\varepsilon_1}, \ldots, \sigma^2_{\varepsilon_4}, \sigma_{\varepsilon_1\varepsilon_2})$. The restricted model (Figure 1a), which corresponds to the researcher's hypothesized model, is defined by fixing $\sigma_{\varepsilon_1\varepsilon_2}$ to zero, and its parameter vector $\boldsymbol{\theta}_r$ is given by $\boldsymbol{\theta}_r = (\tau_1, \ldots, \tau_4, \lambda_1, \ldots, \lambda_4, \sigma^2_{\varepsilon_1}, \ldots, \sigma^2_{\varepsilon_4})$.

The comparison of the models in Figure 1 is one example of a larger class of model comparison problems, which can involve multiple parameters (e.g., θ_{q+1}, θ_{q+2}, …, θ_{q+J}), linear and nonlinear constraints, and modeling approaches beyond traditional factor analysis or SEM (e.g., multilevel models). Our proposed extension of the score test to multiply imputed data readily generalizes to accommodate these contexts. When appropriate, we will indicate the differences between single-parameter and multiple-parameter tests. We chose to use the comparison of the models in Figure 1 as our motivating example both to present some of the more complex statistical concepts in a familiar statistical framework and ground the multiple imputation score test in a context in which researchers routinely use its complete-data counterpart (model modification in SEM).

Multiple Imputation

Multiple imputation dates back nearly 40 years (Rubin, 1987; Rubin, 1996) and is an established and popular method for dealing with missing data. We provide some brief background here and refer readers to the literature for additional information (Enders, 2010;

Graham, 2012; Schafer, 1997; Schafer & Graham, 2002; Schafer & Olsen, 1998; Sinharay, Stern, & Russell, 2001; van Buuren, 2012). Multiple imputation is often described as a three-step procedure. In the first step, the researcher creates many copies of an incomplete data set, each of which is imputed with different estimates of the missing values; for example, Graham, Olchowski, and Gilreath (2007) recommend at least 20 imputed data sets. We use $\mathbf{X}^{(m)},\ m = 1, \ldots, M$, to represent imputed data sets generated from an incomplete data matrix $\mathbf{X}$. Next, the analysis model (e.g., the restricted CFA model from Figure 1a) is fit to each of the filled-in data sets, which gives a set of imputation-specific maximum likelihood estimates $\widehat{\boldsymbol{\theta}}_r^{(m)}$ and an estimate $\widehat{\mathbf{V}}_r^{(m)}$ of the parameter covariance matrix, the diagonal of which contains the complete-data sampling variances (i.e., squared standard errors) for data set m. Finally, Rubin's rules (Rubin, 1987) are applied to the M sets of results, giving a vector of pooled estimates $\widehat{\boldsymbol{\theta}}_r$ and standard errors. As mentioned previously, Lee and Cai (2012) and Enders and Mansolf (2018) discuss model fit statistics for multiply imputed data.

In the context of an SEM application, the imputation step typically employs a saturated model and an iterative Bayesian estimation procedure such as the Gibbs sampler. Because imputations are generated from a saturated model, they are appropriate for a range of nested model comparisons beyond that depicted in Figure 1. The iterative algorithm alternates between two major steps: (a) estimate saturated model parameters, conditional on the current filled-in data set, then (b) update imputations conditional on the current model parameters. Joint modeling (Asparouhov & Muthén, 2010; Schafer, 1997) and fully conditional specification (van Buuren, 2012; Van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006) are the primary frameworks for generating multiple imputations. The joint model approach repeatedly estimates a multivariate model, typically an unrestricted mean vector and covariance matrix. The imputation

step then samples replacement values from a multivariate normal distribution. In contrast, fully conditional specification implements a round robin imputation scheme where each incomplete variable is regressed on all other variables (complete or previously imputed), and the resulting regression model parameters define a univariate distribution of replacement values for each observation. Joint model imputation is available in commercial packages such as SAS and Mplus (Muthén & Muthén, 1998–2017) as well as R packages such as `jomo` (Quartagno & Carpenter, 2018) and `Amelia` (Honaker, King, & Blackwell, 2018), among others. Fully conditional specification is available in commercial software packages such as SPSS and SAS, and it is also available in the popular R package `mice` (van Buuren et al., 2018) and the standalone application Blimp (Enders, Du, & Keller, in press; Enders, Keller, & Levy, 2018; Keller & Enders, 2019). Both approaches can accommodate incomplete binary and ordinal variables via a latent variable (i.e., probit regression) formulation, and the procedures are theoretical equivalent when applied to multivariate normal data (Liu, Gelman, Hill, Su, & Kropko, 2014).

It is important to address a compelling question before proceeding—why imputation? After all, FIML test statistics are well understood, and the estimator is available in virtually every SEM software program. We believe there are often compelling reasons to adopt imputation instead of FIML. First, imputation is arguably more flexible for models that include mixtures of continuous and categorical (nominal and ordinal) variables. Such combinations of metrics are ubiquitous in SEM applications, for example MIMIC models with categorical covariates, multiple-group models with incomplete grouping variables, measurement models that feature discrete questionnaire responses as indicators, and scale scores or item parcels computed from an incomplete set of questionnaire items, among others. Second, emerging evidence suggests that a particular model-based variant of multiple imputation (e.g., fully Bayesian imputation;

substantive model compatible imputation) is superior for models that include interactive or nonlinear effects and random coefficients (Enders et al., in press; Erler, Rizopoulos, Jaddoe, Franco, & Lesaffre, 2019; Erler et al., 2016). Third, the ease with which multiple imputation can facilitate an inclusive analytic strategy (Collins, Schafer, & Kam, 2001) that includes auxiliary variables is a potentially important advantage. Of course, FIML estimation can also incorporate auxiliary variables (Graham, 2003), but it is well known that the saturated correlates model can suffer from convergence problems because it imposes an illogical structure on certain covariance matrices (Savalei & Bentler, 2009). For these and other reasons, we believe that marrying SEM and multiple imputation is often a preferable strategy for handling missing data, in which case it is important to have a full cadre of significance tests that includes the score test.

Statistical Background: Score, Score Vector, and Information Matrix

In the interest of space, this section assumes that readers are already familiar with the maximum likelihood principles needed to construct the complete-data (or FIML) score test. In particular, the building blocks of the test statistic are the score, the score vector and the information matrix. This section describes these quantities in the context of multiple imputation. In general, the imputation-based expressions are the same as those from a complete-data maximum likelihood analysis, the key difference being that they are applied to each filled-in data set. To ensure that our exposition is accessible to the broadest possible readership, we also include online supplemental material (Supplemental Appendix A) that provides a more detailed description and summary of the concepts needed to understand the composition of the score test statistic, and a variety of resources describe maximum likelihood estimation in greater depth (Eliason, 1993; Ferron & Hess, 2007; Silvey, 1975, Chap. 4; Casella & Berger, 2002, Chap. 7; Spanos, 1999, Chap. 13).

To keep the discussion as simple as possible, we describe a single degree of freedom score test that evaluates a constraint on a single parameter. The test readily extends to multiple parameters, and Supplemental Appendix B gives the generalization of the test statistic to multiple parameter constraints. The first ingredients of the modification index are the score and the score vector. In calculus terms, the score for a parameter is the derivative or slope of the log-likelihood surface taken at a particular value of that parameter, and the score vector concatenates the parameter-specific score values into a vector that quantifies the *instantaneous* rate of change in the log-likelihood with respect to each model parameter. For the restricted model in imputation *m*, the score vector $\mathbf{S}_r^{(m)}$ has *q* elements, one for each parameter in $\boldsymbol{\theta}_r^{(m)}$, and each element of the score vector is the first partial derivative of the log-likelihood with respect to the corresponding model parameter

$$\mathbf{S}_r^{(m)} = \mathbf{S}_r\left(\mathbf{X}^{(m)}\middle|\boldsymbol{\theta}_r\right) = \frac{\partial}{\partial \boldsymbol{\theta}_r} LL_r\left(\mathbf{X}^{(m)}\middle|\boldsymbol{\theta}_r\right) \quad (1)$$

where $\frac{\partial}{\partial \boldsymbol{\theta}_r}$ indicates that partial derivatives of the log likelihood $LL_r\left(\mathbf{X}^{(m)}\middle|\boldsymbol{\theta}_r\right)$ are taken with respect to each element of $\boldsymbol{\theta}_r$, and the *m* superscripts indicate that each data set yields a unique score vector. At the maximum likelihood estimates of the restricted model in a given imputed data set *m*, denoted by $\widehat{\boldsymbol{\theta}}_r^{(m)}$, the *q* elements of the score vector, denoted $\widehat{\mathbf{S}}_r^{(m)}$, all equal zero, indicating that the estimates maximize the log-likelihood function.

The score test is used to determine whether freeing a constraint would significantly improve model fit. In the case of the models in Figure 1, this involves estimating the restricted model, then projecting how fit would change if the residual covariance $\sigma_{\varepsilon_1\varepsilon_2}$ were freed during

estimation. Doing so requires an augmented parameter vector $\widehat{\boldsymbol{\theta}}_g^{(m)} = \left(\widehat{\boldsymbol{\theta}}_r^{(m)}, 0\right)$ that contains the restricted-model maximum likelihood estimates $\widehat{\boldsymbol{\theta}}_r^{(m)}$ within each imputed data set and a zero[2] value corresponding to the constrained parameter θ_{q+1}. Similarly, we define a score vector $\hat{\mathbf{S}}_g^{(m)}$ that reflects the gradient of general-model log likelihood taken at the parameter values in $\widehat{\boldsymbol{\theta}}_g^{(m)}$, given by

$$\hat{\mathbf{S}}_g^{(m)} = \frac{\partial}{\partial \boldsymbol{\theta}_g} LL_g\left(\mathbf{X}^{(m)} \middle| \widehat{\boldsymbol{\theta}}_g^{(m)} = \left(\widehat{\boldsymbol{\theta}}_r^{(m)}, 0\right)\right) = \left(\mathbf{0}_q, \hat{S}_{g,q+1}^{(m)}\right) \tag{2}$$

where $\hat{S}_{g,q+1}^{(m)}$ is the slope of the log likelihood that results from constraining parameter θ_{q+1} (e.g., the residual covariance) to zero. Even when the restricted model is true in the population, the additional element of the score vector is unlikely equal to zero in practice because it too is an estimate subject to sampling error. A positive score (i.e., slope) indicates that increasing the value of the θ_{q+1} would increase the log-likelihood, whereas a negative score indicates that decreasing the parameter's value by a small amount would increase the log-likelihood.

The score vector from Equation 2 can be viewed as measuring the discrepancy in fit between the general and restricted models. The score test we are proposing uses the information matrix to standardize this discrepancy into a test statistic. Like the score vector, the information matrix is calculated by differentiating the log-likelihood function with respect to the model parameters. In calculus terms, the information matrix for imputation set *m* is the matrix of second derivatives of the log-likelihood surface (i.e., negative of the Hessian), as follows.

[2] More generally, the constrained parameter in the augmented parameter vector could also be fixed to any nonzero value (e.g., a factor correlation fixed to 1 to test redundancy), and the score test procedure would still apply.

$$\mathbf{I}_r^{(m)} = -\frac{\partial^2}{\partial \boldsymbol{\theta}_r \partial (\boldsymbol{\theta}_r)^T} LL_r\left(\mathbf{X}^{(m)} \middle| \boldsymbol{\theta}_r^{(m)}\right). \quad (3)$$

The information matrix captures how the elements of the score vector (i.e., gradients or slopes) change as a function of changes in the model parameters. Visually, values on the diagonal of the information matrix quantify the "peaked-ness" of the log likelihood near its maximum (e.g., if slopes change rapidly, the surface is peaked and precision is high). Off-diagonal element (*i,j*) of the information matrix quantifies the change in the *i*th element of the score vector as a function of changes in the *j*th parameter (e.g., the degree to which changes in one parameter covary with changes in another).

The information matrix can be used as a metric by which to judge the magnitude of (i.e., standardize) the score vector, as large elements of the information matrix indicate that the score vector is sensitive to changes in the model parameters near $\widehat{\boldsymbol{\theta}}_r^{(m)}$ (i.e., the log likelihood is very peaked near its maximum). Many readers are familiar with the fact that inverting the information matrix and substituting the maximum likelihood estimates gives the parameter covariance matrix, the diagonal of which contains the complete-data sampling variances (i.e., squared standard errors) for data set *m*:

$$\mathbf{I}_r^{-1(m)} = \widehat{\mathbf{V}}_r^{(m)} = \begin{bmatrix} var\left(\hat{\theta}_1^{(m)}\right) & cov\left(\hat{\theta}_1^{(m)}, \hat{\theta}_2^{(m)}\right) & \cdots & cov\left(\hat{\theta}_1^{(m)}, \hat{\theta}_2^{(m)}\right) \\ cov\left(\hat{\theta}_2^{(m)}, \hat{\theta}_1^{(m)}\right) & var\left(\hat{\theta}_2^{(m)}\right) & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ cov\left(\hat{\theta}_q, \hat{\theta}_1^{(m)}\right) & \cdots & \cdots & var\left(\hat{\theta}_q^{(m)}\right) \end{bmatrix}. \quad (4)$$

From this relationship, we can see that large elements of the information matrix indicate a high degree of precision for the maximum likelihood estimates, thus resulting in small standard errors.

Note that SEM packages typically offer at least three methods for estimating the information matrix: the *first-order*, *observed*, and *expected* information matrices. Details on the calculation of these matrices and on their use in structural equation modeling can be found in Appendix A of the online supplement, and Savalei (2010a) gives a detailed and accessible description of these information matrices[3]. Very briefly, the scalar formulas for the elements of the information matrix feature deviation score sums that capture the differences between the observed data and the model-implied means. If the data are normally distributed, the expected information is simpler to compute because it replaces these sums with their expectations (i.e., zeros), whereas the observed information computes the sums from the data. When applied to FIML estimation, the observed information is preferred because it accommodates an MAR mechanism, whereas the expected information requires the stricter MCAR assumption (Kenward & Molenberghs, 1998). The first-order information matrix is calculated as the covariance matrix of first derivatives of each observation's log-likelihood function and is asymptotically equivalent to the expected and observed information matrices if the distributional assumptions of the model hold (Greene, 2012, pp. 521–522), although it tends to perform worse than the other information matrices in practice (Maydeu-Olivares, 2017). Either expected or observed information is appropriate for imputed data, and the simulations presented later in the manuscript provide a comparison of the three information matrices.

Multiple Imputation Score Test

[3] Most estimation software packages, and most statistical treatments of the topic, divide the "total" information matrix by *n* to yield "unit" information, which we have omitted for simplicity.

Having established its core building blocks, we now propose a multiple imputation score test. The construction of the test statistic is analogous to that of the popular Wald test (commonly referred to as the D_1 statistic) for multiply imputed data (Li, Raghunathan, & Rubin, 1991; Rubin, 1987; Schafer, 1997; van Buuren, 2012). To provide a comparison, the multiple imputation Wald statistic is

$$D_1 = \frac{\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)'\left((1+\bar{r})\overline{\mathbf{V}}\right)^{-1}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)}{k}. \quad (5)$$

The Wald test can be viewed as the sum of squared differences between a vector[4] of pooled estimates $\widehat{\boldsymbol{\theta}}$ and their corresponding hypothesized values $\boldsymbol{\theta}_0$ (e.g., a zero vector) standardized by a pooled estimate of the parameter covariance matrix $(1+\bar{r})\overline{\mathbf{V}}$, where $\bar{r}$ is the average relative increase in variance and $\overline{\mathbf{V}}$ is the pooled within-sample variance-covariance matrix of the model parameters ($\overline{\mathbf{V}} = \frac{\mathbf{1}}{\boldsymbol{m}}\sum \widehat{\mathbf{V}}_r^{(m)}$). Finally, dividing by k, the number of parameters in $\widehat{\boldsymbol{\theta}}$, rescales the test statistic from a χ^2 to F distribution. Imputation-based significance tests are quite different from those of FIML because they require a between-imputation component that captures variation in the test across data sets (Li et al., 1991; Meng & Rubin, 1992; Reiter & Raghunathan, 2007; Rubin, 1987). The average relative increase in variance ($0 \geq \bar{r} \geq 1$) serves this role by proportionally increasing the average parameter covariance matrix $\overline{\mathbf{V}}$ to incorporate lack of precision due to missing data. The score test below is comprised of analogous components.

[4] Although the preceding discussion of the score test referred to an augmented vector containing only a single constrained parameter estimate, the score test also generalizes to simultaneously testing a set of parameters (Supplemental Appendix B).

First, the nonzero element (or elements, in the multiparameter case) of the score vector from Equation 2 is pooled across imputations to yield an average discrepancy in fit between the general and restricted models.

$$\bar{S}_{g,q+1} = M^{-1} \sum_{m=1}^{M} S_{g,q+1}^{(m)} \tag{6}$$

This averaging process is analogous to the sum of squared differences between the estimates and hypothesized values in the Wald test; $\bar{S}_{g,q+1}$ quantifies the discrepancy between zero (the value of the constrained parameter in the restricted model) and the projected maximum likelihood estimate of the constrained parameter in the general model; this difference is quantified by the average slope of the log likelihood function. The pooled score value $\bar{S}_{g,q+1}$ can be interpreted in frequentist terms as a point estimate of the fixed population score value, and it can also be interpreted as the mean of the observed-data posterior distribution of the score in the Bayesian framework (Little & Rubin, 2002, pp. 210-211; Rubin, 1987).

As mentioned previously, the pooled discrepancy measure is standardized using the curvature of the log likelihood function from the information matrix. If the null hypothesis is true and the data were generated according to the restricted model, $S_{g,q+1}^{(m)}$ is asymptotically normal with a mean of zero. To determine the sampling variance of $S_{g,q+1}$, we must partition the information matrix $\mathbf{I}_g^{(m)}$ to isolate the term corresponding to the parameter of interest

$$\mathbf{I}_g^{(m)} = \begin{bmatrix} \mathbf{P}^{(m)} & \mathbf{Q}^{(m)} \\ \left(\mathbf{Q}^{(m)}\right)' & r^{(m)} \end{bmatrix} \tag{7}$$

where $\mathbf{P}^{(m)}$ is the $q \times q$ submatrix of $\mathbf{I}_g^{(m)}$ corresponding to the free parameters in the restricted model, $\mathbf{Q}^{(m)}$ is the $1 \times q$ submatrix of $\mathbf{I}_g^{(m)}$ containing second partial derivatives of the log-likelihood with respect to the fixed and free parameters, and $r^{(m)}$ is the diagonal element of $\mathbf{I}_g^{(m)}$ corresponding to the fixed parameter θ_{q+1} (e.g., the residual covariance in Figure 1b). Following the complete-data literature (Sörbom, 1989; Saris, Satorra, & Sörbom, 1987; Rao, 1948) the sampling variance of $S_{g,q+1}^{(m)}$ from data set *m* is as follows.

$$v^{(m)} = r^{(m)} - \mathbf{Q}^{(m)} \left(\mathbf{P}^{(m)}\right)^{-1} \left(\mathbf{Q}^{(m)}\right)' \tag{8}$$

Roughly speaking, $v^{(m)}$ can be viewed as the sampling variance in the score vector that remains after subtracting out its covariation due to other parameters.

The *within-imputation* score variance is computed by averaging the *M* estimates of v as follows.

$$v_W = M^{-1} \sum_{m=1}^{M} v^{(m)} \tag{9}$$

The within-imputation score variance v_W is analogous to the within-imputation variance of the parameter estimates in multiple imputation inference in the sense that it quantifies the sampling variance in the score value that would be expected if the data were complete, much like the average of the squared standard errors on the diagonal of $\overline{\mathbf{V}}$ in the Wald test. As such, a score test

statistic based on v_W will be positively biased (i.e., reflect too much precision) unless there is no missing data.

Following the logic of the Wald test, a multiple imputation score test must augment v_W with a second between-imputation component that quantifies the added sampling variability in $\bar{S}_{g,q+1}$ due to missing data. This *between-imputation score variance* is calculated by treating the set of imputation-specific score values $S_{g,q+1}^{(m)}$ as an i.i.d. random sample and calculating the variance of the *M* score values around their pooled value as follows.

$$v_B = (M-1)^{-1} \sum_{m=1}^{M} \left(S_{g,q+1}^{(m)} - \bar{S}_{g,q+1} \right)^2 \tag{10}$$

This quantity represents the added uncertainty in the score value due to missing data and is analogous to the between-imputation variance of the parameter estimates (i.e., the variance in the *M* estimates around their means) in multiple imputation inference.

Combining the within- and between-imputation information components v_W and v_B yields the *total score variance* v_T.

$$v_T = v_W + \left(1 + \frac{1}{M}\right) v_B \tag{11}$$

In frequentist terms, v_T estimates the total variability of the observed-data score value across repeated samples. In Bayesian terms, v_T represents the total posterior variance of the score value, calculated as the sum of a within-imputation component v_W based on complete data and a between-imputation component $\left(1 + \frac{1}{M}\right) v_B$ which accounts for missing data uncertainty.

Finally, the total sampling variance v_T is used to standardize the squared pooled score value and construct the multiple imputation score test statistic below.

$$T_{MI-score} = \frac{\left(\bar{S}_{g,q+1}\right)^2}{v_T} \tag{12}$$

Determining a sampling distribution for $T_{MI-score}$ is complicated by the fact that v_B is an estimate of the between-imputation variance of $S_{g,q+1}^{(m)}$ based on a finite number of imputations. With an infinite number of imputations, we can treat v_B as fixed, in which case $T_{MI-score}$ asymptotically follows a χ^2 distribution with a single degree of freedom (Rubin, 1987). With a finite number of imputations *M*, the proper reference distribution is an *F* distribution with a single numerator degree of freedom and denominator degrees of freedom

$$v = (M-1)\left(1 + \frac{v_W}{v_B + \frac{v_B}{M}}\right)^2. \tag{13}$$

This calculation is based on the degrees of freedom calculation for the multiple imputation Wald test (Rubin, 1987). Alternate definitions of the degrees of freedom have been proposed in the literature, but we do not consider these here (Barnard & Rubin, 1999; Li et al., 1991; Reiter, 2007).

Like the pooled Wald test statistic (D_1) in Equation 5, our example of testing a single parameter in Equation 12 is naturally extended to the case of a multiparameter test by augmenting the score vector (in Equation 2) and information matrix (in Equation 7) with $k > 1$

constrained or fixed parameters. With sufficiently many imputations, the numerator of Equation 5 is approximately $\chi^2(k)$ distributed, which is scaled by $1/k$ (the denominator of Equation 5) to yield an approximate F distribution to account for a finite number of imputations. Details about the multiparameter score test can be found in supplemental Appendix B.

Expected Parameter Change

Saris et al (1987) noted that large modification indices do not necessarily result in comparably large changes in parameter estimates when the corresponding parameters are freed during estimation. In other words, although the modification index tests the statistical significance of parameters, it does not quantify the magnitude of the constrained estimate. To account for this shortcoming, Saris et al. introduced a statistic called the expected parameter change (EPC) which quantifies the expected change in the parameter estimate that would result from freeing that parameter during estimation. For the model in Figure 1, the EPC estimates the residual covariance $\sigma_{\varepsilon_1\varepsilon_2}$ that would result from estimating the general model. These EPC values are routinely printed alongside modification indices in structural equation modeling software (Muthén & Muthén, 1998–2017; Rosseel, 2012).

This EPC value, like the score test, is calculated after estimating the restricted model, and is a function of the test statistic (i.e., modification index) and the score value. The complete-data EPC is as follows.

$$EPC = \frac{T_{score}}{S_{g,q+1}} \tag{14}$$

As explained previously, the nonzero element of the score vector $S_{g,q+1}$ is the expected change in the log-likelihood for a one-unit change in the parameter of interest. Thus, its reciprocal is the

expected change in the parameter for a one-unit change in the log-likelihood. Multiplying this reciprocal by the change in the log-likelihood that results from freeing the parameter, quantified by T_{score}, yields the EPC. Because the general model is not estimated, EPC values will differ from the estimates that would result from freeing each residual covariance. If the models do not differ substantially in terms of fit, the differences between the EPC values and the true parameter estimates will be small. However, if model fit is poor, the score test statistic and resulting EPC values can be biased (Saris et al., 1989).

In the context of imputation, we defined the pooled EPC as the average of the EPC values across the *M* imputations

$$\text{EPC}_{MI} = M^{-1} \sum_{m=1}^{M} \frac{\left((S_{g,q+1}^{(m)})^2 / v^{(m)} \right)}{S_{g,q+1}^{(m)}} \tag{15}$$

where $S_{g,q+1}^{(m)}$ is element $q + 1$ of the imputation-specific score vector $\mathbf{S}_g^{(m)}$ and $v^{(m)}$ is sampling variance of the score value in data set *m*, as defined in Equation 9. Note that, unlike its complete-data counterpart, EPC_{MI} is not a function of the score test statistic $T_{MI-score}$. Rather, it is estimated by averaging the imputation-specific EPC values, each of which is calculated using its corresponding score variance $v^{(m)}$.

Simulation Study

We conducted a simulation study to assess the performance of the multiple imputation score test in terms of Type I error rate and statistical power. We compared the multiple imputation score test to the FIML score test and to the score test from complete-data maximum likelihood (i.e., before missingness was imposed on the generated data). As mentioned

previously, the score test is asymptotically equivalent to the Wald and likelihood ratio statistics as general hypothesis tests. In the interest of space, we limit our presentation to the score tests, but Figures S1-S4 and Tables S1 and S2 in the online supplemental materials shows a broader comparison that includes the Wald and likelihood ratio statistics.

Population Models

We based the population model on a classic simulation study (Chou & Bentler, 1990) that compared the relative performance of the Wald, likelihood ratio, and score tests in complete data. The population model used for the simulation study is displayed in Figure 2. In the measurement portion of the model, each of four latent variables F_1-F_4 is measured using three indicator variables with factor loadings of 1 (X_1, X_4, X_7, X_{10}) or 0.8 (X_2, X_3, X_5, X_6, X_8, X_9, X_{11}, X_{12}). Residual variances for observed variables were set to one third of the squared factor loading for each observed variable, yielding item-level reliabilities of .75.

In the structural portion of the model, F_1 is correlated with F_2 (φ_{12}), F_3 is predicted by F_1 (β_{13}) and F_2 (β_{23}), and F_4 is predicted by F_1 (β_{14}), F_2 (β_{24}), and F_3 (β_{34}). Population values for the correlation φ_{12} and the regression coefficients β_{13} and β_{23} were fixed to values such that the correlations between F_1, F_2, and F_3 were all equal to .3 ($\varphi_{12} = .3$; $\beta_{13} = \beta_{13} \approx .231$). Population values for the remaining structural parameters β_{14}, β_{24}, and β_{34} were determined by the conditions of the corresponding simulation study, to be described below. The means of all observed and latent variables were set to zero in the population model.

Simulation Conditions

In the first simulation study, β_{14}, β_{24}, and β_{34} were set to zero in the population model to examine the Type I error rate, which was assessed separately for one- (β_{14}), two- (β_{14} and β_{24}), and three- (β_{14}, β_{24}, and β_{34}) parameter Wald, score, and likelihood ratio tests. Sample sizes used

to evaluate Type I error rates were 100, 200, 400, and 800, and tests were evaluated with 0% (complete data), 10%, 20%, and 30% missing data, yielding a 4 (sample size) × 4 (missing data rate) × 3 (number of parameters) design.

In the second simulation study, each of the one-, two-, and three-parameter tests above was evaluated under varying effect size conditions to examine statistical power and assess the accuracy of EPC estimates. To yield interpretable power estimates, we manipulated the population values of β_{14}, β_{24}, and β_{34} to yield (multiple) R^2 values of .02, .13, and .26, corresponding to Cohen's (1988) small, medium, and large effect sizes, for the latent regression predicting F_4. All tested structural parameters β_{14}, β_{24}, and β_{34} were identical for each condition in the population model; see Table 1 for the specific parameter values used. For the one- and two- parameter tests, any untested parameters were set to zero in the population model so that test statistics for truly nonzero parameters would not be inflated due to invalid constraints on untested parameters (i.e., higher power in conditions with inflated Type I error rates; Bollen, 1989; Byron, 1972). Sample sizes used to evaluate power were 50, 100, and 200, and tests were evaluated with 0% (complete data), 10%, 20%, and 30% missing data, yielding a 3 (sample size) × 4 (missing data rate) × 3 (number of parameters) × 3 (R^2 effect size) design. EPC values were calculated for the regression coefficient β_{14} in the one-parameter test conditions for all information matrices in complete and incomplete data and for FIML estimation in incomplete data.

Data Simulation and Imputation

Data were simulated according to a multivariate normal distribution with all means equal to zero and covariance matrix equal to the model-implied covariance matrix of the population model, calculated according to the corresponding simulation condition. After simulating 1000

data sets per condition, data were deleted according to a missing-at-random (MAR) model, where all indicators for F_1 predicted missingness on all indicators for F_2 and all indicators for F_3 predicted missingness on all indicators for F_4. Missingness was determined according to a logistic regression model with a pseudo-R^2 value of .5 (McKelvey & Zavoina, 1975), with parameters of β_0 = -3.22, β_1=1.81 corresponding to a 10% missing data rate, parameters of β_0 = -2.10, β_1=1.81 corresponding to a 20% missing data rate, and parameters of β_0 = -1.30, β_1=1.81 corresponding to a 30% missing data rate. R (R Core Team, 2018) was used to perform the data generation using the `mvtnorm` package (Genz et al., 2018). Blimp (Enders et al., in press; Enders et al., 2018; Keller & Enders, 2019) was used to impute missing values using fully conditional imputation (van Buuren, 2012; van Buuren et al., 2006). Because all manifest variables in the simulation were continuous, imputations were generated from a normal distribution, conditional on the observed data. In all conditions, 20 imputations were used, with burn-in and thinning intervals determined after examining potential scale reduction factor diagnostics (Gelman et al., 2014; Gelman & Rubin, 1992).

Statistical Analyses and Tests

We used the `sem()` function in the R package `lavaan` (Rosseel, 2012) to fit the restricted and general models implied by the path diagram in Figure 2 to the generated data, both before missingness was imposed on the complete data (using traditional maximum likelihood estimation) and after imposing missingness (using FIML). After imputing the missing data 20 times, we used the `sem.mi()` function in the R package `semTools` (Jorgensen et al., 2019) to fit the models to each imputed data set

The restricted and general models had correctly specified measurement models (see population details above), identifying each factor's variance by fixing its first indicator's factor

loading to one (corresponding to its population value) and identifying factor means by fixing them to zero. In the general model, all structural paths in Figure 2 were freely estimated, whereas one, two, or three of the nonzero dashed paths (depending on the condition) were fixed to zero in the restricted model. Analyses of multiple imputations were treated as converged if the general or restricted model converged for at least one imputed data set(s)[5].

From the restricted model, a score test statistic was calculated to simultaneously test whether the one, two, or three (depending on the condition) nonzero dashed paths in Figure 2—which were fixed to zero—should be freed. Score tests for complete-data and FIML were conducted using the `lavTestScore()` function in `lavaan`, and using the `lavTestScore.mi()` function in `semTools`, for multiple imputations. The pooled, augmented information matrix with which a pooled score test is calculated (as described above) used any imputations for which the model converged and *SE*s (and by implication, the information matrix) could be calculated. Additionally from the restricted model alone, EPCs for individual fixed-to-zero parameters are available from the `modificationIndices()` function in `lavaan`, and from the `modificationIndices.mi()` function in `semTools` for multiple imputations, both of which return 1-*df* score tests (i.e., modification indices) only for individual parameters, not multiparameter tests[6].

As mentioned previously, SEM packages usually offer at least three methods for estimating the information matrix: the *first-order*, *observed*, and *expected* information matrices.

[5] The semTools package prints a message to make users aware of how many imputations for which a model converged, so the user can decide whether to generate additional or replacement imputations or to try different starting values for particular imputations.

[6] The lavTestScore() and lavTestScore.mi() functions also optionally provide EPCs for all user-specified (both fixed and free) parameters, but they are the expected changes on the condition that all parameters in the test are freed. This can be quite useful information (e.g., Oberski, 2014), but it differs from the most common use of EPCs (i.e., in tandem with 1-*df* modification indices; Saris et al., 2009; Whittaker, 2012). Thus, the current study only investigated EPCs associated with 1-*df* modification indices.

Details on the calculation of these matrices and on their use in structural equation modeling can be found in Appendix A of the online supplemental material and in Savalei (2010a). We used observed information exclusively with FIML estimation because it provides important theoretical advantages in this context (Kenward & Molenberghs, 1998). Applied to imputed data, the observed information does not have an inherent theoretical advantage over other estimates of information as it does with FIML. To assess the performance of different information matrices – all of which are available in popular software (Muthén & Muthén, 1998-2017; Rosseel, 2012) and thus could be used in practice – we evaluated the performance of the multiple imputation score test with observed, expected, and first-order information matrices. Although researchers with different goals could validly justify different arbitrary thresholds for qualifying Type I error rates as substantially inflated, we define an acceptable range for Type I error rates as $2.5\% \leq \alpha \leq 7.5\%$ and Type I error rates $7.5\% < \alpha \leq 10\%$ and $\alpha > 10\%$ as moderate and large inflation, respectively (Savalei, 2010a; Savalei, 2010b).

Simulation Results

As described above, the score test can be evaluated as an approximate χ^2 or F statistic. As might be expected with a large number of imputations ($m = 20$), we did not observe meaningful differences between p-values calculated using the two reference distributions. Because the FIML and complete-data score tests use a χ^2 reference distribution, we limit our attention to χ^2 tests here to improve comparability of the simulation results. Although we limit the subsequent presentation to the score tests, Figures S1-S4 and Tables S1 and S2 in the online supplement give a broader comparison that includes Wald and likelihood ratio statistics. In some cases, when the null hypothesis for the 1-, 2-, or 3-parameter test was false (power and EPC simulations), certain test statistics could not be calculated. Specifically, when sample size was low ($N = 50$) and

missing data rate was high (30%), up to 12.2% of multiple imputation likelihood ratio tests could not be calculated, with higher failure rates for multi-parameter tests. In six replications, either the unrestricted (four replications) or restricted (two replications) model failed to converge under FIML estimation. The two replications for which the restricted model failed to converge, both of which occurred at the smallest sample size ($N = 50$) and highest missing data rate (30%) in the Power simulation, were excluded from Figures 5 and 6. All replications for which all score tests could be calculated were included in the results presented in Figures 4-5.

Type I error

Like its FIML counterpart, the multiple imputation score test can be computed using three different information matrices, although it is unclear whether the usual recommendation to use observed information (Kenward & Molenberghs, 1998; Savalei, 2010a) also applies to multiple imputation. Figure 3 presents empirical Type I error rates for the three versions of the test. As sample size increased to $N = 800$, all three information matrices produced acceptable Type I error rates close to the nominal $\alpha = .05$ level. However, differences appeared at small sample sizes. The multiple imputation score test achieved its best calibration when computed using the expected information, in which case the empirical Type I error rates were close to the nominal $\alpha = .05$ under almost all conditions. The notable exception was a slightly lower Type I error rate (approximately .03) for the 3-parameter test with 30% missing data and a sample size of 100. The expected information should perform well with normally distributed data (Browne, 1974, 1984), but there was no clear reason to prefer this method a priori.

Among the remaining two options, the test statistic with observed information was the next best calibrated, although it exhibited inflated Type I error rates in many conditions. This is consistent with prior research showing inflated Type I error rates of the likelihood ratio test using

FIML[7] even with samples as large as $N = 400$ (Savalei & Bentler, 2009), especially in models that estimated more parameters (i.e., included more auxiliary variables). The test based on first-order information, however, exhibited the largest deviations from the nominal error rate. Decreasing the sample size, increasing the missing data rate, and increasing the number of parameters being evaluated all diminished performance, and these factors exerted a much larger impact on the first-order information tests than on the other tests. While tests based on observed information were generally well-calibrated in all but the smallest sample size condition ($N = 100$), first-order tests were generally only well-calibrated in the largest sample size condition ($N = 800$). Based on these results, we restrict the remainder of our results to the score test with expected information, and we compare its performance to the FIML test with observed information. We also include the complete-data score test with expected information as a comparison (also the best among the complete-data options; see Table S1 in the online supplement).

Figure 4 shows empirical Type I error rates for the complete-data, multiple-imputation, and FIML score tests. In the smallest sample size condition ($N = 100$), the multiple imputation score test had Type I error rates lower than 5% when calculated with 30% missing data. Within these conditions ($N = 100$, 30% missing data), the Type I error rate of the multiple imputation score test decreased as the number of parameters increased, reaching a minimum of $\alpha = 2.3\%$ with 3 parameters, 30% missing data, and $N = 100$. At missing data rates of 10% and 20%, Type I error rates for the multiple-imputation score test were within acceptable ranges ($2.5\% \leq \alpha \leq 7.5\%$).

[7] The likelihood ratio test also exhibits inflated Type I errors with complete data, and there have been several proposed robust corrections based on aspects of sample size and model size (Nevitt & Hancock, 2004).

The FIML score test, in contrast, exhibited moderate ($7.5\% < \alpha \leq 10\%$) to large ($\alpha > 10\%$) Type I error inflation in many conditions; this inflation increased as sample size decreased, the missing data rate increased, and the number of parameters increased. At $N = 100$, the empirical Type I error rate of the FIML score test was only within acceptable ranges ($2.5\% \leq \alpha \leq 7.5\%$) when used for a single-parameter test with 10% or 20% missing data. The 3-parameter FIML score test with 30% missing data and a sample size of 100 exhibited large Type I error inflation ($\alpha > 10\%$), and the comparable condition with $N = 200$ also exhibited moderate Type I error inflation ($7.5\% < \alpha \leq 10\%$). In contrast, the multiple imputation score test was well-calibrated under these same conditions, except for the aforementioned Type I error deflation when 3 parameters were tested with 30% missing data and $N = 100$. In all other conditions with $N = 100$, the FIML test had moderate Type I error inflation ($7.5\% < \alpha \leq 10\%$), while the multiple imputation score test was well-calibrated. When sample size was large ($N = 400$ or 800), both the multiple imputation and FIML score tests were well-calibrated, while the multiple imputation score test had superior Type I error control to the FIML score test when sample size was small ($N = 100$ or 200). Lastly, the complete-data score test was well-calibrated across conditions and deviated very little from $\alpha = .05$ in any condition.

Power

Figure 5 shows empirical power estimates for the single-parameter multiple imputation, FIML, and complete-data score test statistics. Power differences among the score tests were relatively unaffected by the number of parameters tested, so we limit the presentation to the single-parameter case to highlight the main trends. Table S2 in the online supplement gives the power estimates for the two- and three-parameter tests.

Unsurprisingly, power for the single-parameter score tests increased as the sample size and effect size increased, and power for the incomplete-data tests (multiple imputation and FIML) decreased as the missing data rate increased. At the largest sample sizes considered in the power simulation ($N = 100$ and $N = 200$), there was generally no meaningful difference between the three tests. However, large differences appeared at the smallest sample size condition ($N = 50$), with multiple imputation score test exhibiting the lowest power. For example, in the medium effect size condition with $N = 50$, the power of the imputation-based test was approximately half that of FIML. Presumably, the power differential reflects the fact that multiple imputation involves two stages of estimation (i.e., the first stage employs a saturated model for the purposes of imputation, and the second stage fits the SEM to the filled-in data) whereas maximum likelihood involves one. Similar power differences have been observed for the direct FIML and two-stage FIML estimators (Savalei & Bentler, 2009) and for the multiple imputation and FIML likelihood ratio tests (Enders & Mansolf, 2018).

Although we would expect the complete-data score test to always yield highest power, the incomplete-data FIML test paradoxically had higher power in some situations. This apparent advantage may be related to the FIML score test's inflated Type I error rate, which may indicate a greater tendency to yield low *p*-values in general. As sample size increased, the multiple imputation and FIML score tests approached the upper limit on power (100%), diminishing the apparent differences between the power of the two tests. In those conditions where power differed between the tests, most notably when sample size was small ($N = 50$), the power advantage of the FIML score test over the multiple imputation score test increased with missing data rate and effect size.

Expected parameter change

Figure 6 contains median relative bias for multiple imputation, FIML, and complete-data EPC. Median relative bias was calculated by subtracting the estimated EPC values in each condition from their population value ($\beta \approx .141$ for $R^2 = .02$; $\beta \approx .360$ for $R^2 = .13$; $\beta \approx .510$ for $R^2 = .26$), dividing by the population value, and calculating the median across replications:

$$Median\ relative\ bias = MEDIAN\left(\frac{EPC-\beta}{\beta}\right). \quad (23)$$

We chose to report median relative bias for EPC to account for many severe outliers in FIML EPC. These severe outliers arose primarily from the $R^2 = .26$ condition.

The multiple imputation EPC was unbiased in almost all conditions, with slight downward bias when the missing data rate was high and sample size was low. Complete-data EPC was also unbiased in almost all conditions, with similar but smaller biases occurring in the same conditions as multiple imputation EPC. In contrast, FIML EPC was severely positively biased for all effect sizes except $R^2 = .02$ ($\beta \approx .141$), for which median relative bias was within an acceptable range. The presence of outliers and bias in EPCs using FIML may be due to the tendency for the unrestricted model's observed information matrix to be non-positive-definite or otherwise poorly behaved, a phenomenon which has been documented elsewhere (e.g., Freedman, 2007; Morgan, Palmer, & Ridout, 2007).

Real Data Example

The multiple imputation score test is relatively straightforward to implement in software and is now available in the R package `semTools` (Jorgensen et al., 2019). To illustrate its application, we used a publicly available online data set containing responses to the Rosenberg Self-Esteem Scale (RSES) (Rosenberg, 1965). These data were downloaded from the

http://personality-testing.info/_rawdata/ webpage. According the website, “Users were informed at the beginning of the test that there [sic] answers would be used for research and were asked to confirm that their answers were accurate and suitable for research upon completion (those that did not have been removed from these datasets).” The item content for the RSES is presented in Table 2. The original data contained 47,974 individual response vectors, and we selected 1000 cases for the example. The data are available from the first author upon request.

To better control the missing data mechanism, we restricted the data to contain only cases between 18 and 64 years of age with complete data for the Rosenberg items, age, and gender (N = 34660), then selected 1000 cases at random to analyze, which constituted the *complete data set*. We then deleted item response data for Items 5, 8, 9, and 10 according to a linear regression model where missingness was predicted by age, yielding the *incomplete data set*. Logistic regression parameters were selected to yield approximately 15% missing data and McKelvey and Zavoina pseudo-R^2 of .5 between age and missingness (standardized $b_0 = -2.6$, $b_1 = 1.815$).

We used fully conditional specification (i.e., the “mice” algorithm) in Blimp (available at www.appliedmissingdata.com) to generate 50 data sets. To maintain the metric of the original variables, discrete imputes were generated from an ordinal probit model (Carpenter & Kenward, 2013). We then fit a one-factor measurement model (analogous to Figure 1a) that featured the 10 RSES items as indicators to each imputed data set; we fit this model using maximum likelihood estimation, treating the observed data as continuous, to illustrate the multiple imputation score test and EPC in the maximum likelihood context in which we derived them above. We used the `semTools` and `lavaan` packages to pool the estimates, calculate fit measures, implement the score test, and compute EPCs. The basic R syntax for the analysis is given in the Appendix to illustrate how the package can be used (more extensive annotated syntax is provided in Section C

of the online supplement). Consistent with previous findings that a one-factor model does not adequately describe the RSES (Huang & Dong, 2012), the pooled likelihood ratio test statistic and pooled fit indices (Enders & Mansolf, 2018; Meng & Rubin, 1992) indicated poor fit: $\chi^2(35) = 600.90$, CFI = .879, TLI = 0.845, RMSEA = 0.134. Next, we used the score tests to identify residual covariances that might improve model fit. Table 3 contains the resulting test statistics and corresponding EPCs. Consistent with the findings of Reise, Kim, Mansolf, and Widaman (2016), the three largest modification indices and EPCs were for Items 9 and 10, Items 1 and 2, and Items 6 and 7. A detailed tutorial on the use of the `semTools` package for calculating the multiple imputation score test and EPC values can be found in online supplemental materials.

Discussion

Structural equation modeling (SEM) applications routinely employ a trilogy of significance tests that includes the likelihood ratio test, Wald test, and score test or modification index. Researchers use these tests to assess global model fit, evaluate whether individual estimates differ from zero, and identify potential sources of local misfit, respectively. The FIML versions of these tests have received considerable attention in the methodology literature (Kenward & Molenberghs, 1998; Savalei, 2010a, 2010b; Savalei & Bentler, 2009; Savalei & Yuan, 2009; Yuan et al., 2014; Yuan & Bentler, 2000, 2010; Yuan & Savalei, 2014; Yuan & Zhang, 2012). However, much less is known about test statistics for multiply imputed data. In particular, methodologists have yet to develop a general score test for multiply imputed data, much less one that can serve as a modification index of local misfit in SEM analyses. As such, the goal of this paper was to outline a new score test procedure and use Monte Carlo computer simulations to evaluate its performance.

As sample size increased, the multiple imputation and FIML score tests converged toward optimal Type I error rates and maximum statistical power, although at smaller sample sizes and/or high missing data rates, the FIML score test had inflated Type I error rates, while the multiple imputation score test did not. On the other hand, the multiple imputation score test generally had substantially lower power than the FIML score test when the sample size was small ($N = 50$). Presumably, this is because multiple imputation invokes two stages of estimation whereas FIML invokes only one. Specifically, the first stage estimates a saturated model in order to generate imputations, then the second stages fits the restricted model to the filled-in data. In contrast, FIML can be viewed as "imputing" the data directly from the restricted model, thus eliminating the initial stage of missing data handling. Similar power differences have been observed for the direct FIML and two-stage FIML estimators for missing data (Savalei & Bentler, 2009) and for the multiple imputation and FIML likelihood ratio tests (Enders & Mansolf, 2018). The stark power difference effectively disappeared with a sample size of $N = 100$, but the potential for such large differences suggests that future studies should thoroughly probe the intersection of sample size and effect size.

EPC estimates based on the multiple imputation score test were considerably more accurate than the corresponding FIML EPCs, particularly when effect size for the omitted parameter was moderate or large. Unlike with the test statistics themselves, EPC bias for FIML increased, rather than decreased, with a larger sample size. Taken as a whole, our results suggest that the imputation-based score test is comparable if not superior to that of FIML, at least in the limited conditions we investigated here. At least in part, it seems that using expected versus observed information might play a role in producing this difference; the latter is recommended for FIML, whereas imputation can accommodate either. Importantly, the expected information

requires the multivariate normality assumption, so we do not feel comfortable concluding that the imputation-based test will outperform FIML in general. Future research should attempt to clarify these issues.

One central issue is how to compute the score test in SEM software. Although SEM software packages naturally produce all of the ingredients (e.g., the score vector, information matrix), combining the component parts requires some effort. To facilitate its application, the R package `semTools` now implements the imputation-based score test, and it allows users to choose[8] an information matrix. Our simulation results clearly favor the expected information, but additional research is needed to determine whether this recommendation generalizes to a broader array of conditions. In the interim, it may be wise to conduct a sensitivity analysis to determine whether one's conclusions about local misfit are stable across computational options. Based on complete-data research and the results presented here, we would not expect the first-order information matrix to perform well (see Maydeu-Olivares, 2017), but `semTools` nevertheless offers this option.

It is important to reinforce previous warnings about the data-driven use of modification indices in sequential specification searches (Bollen, 1989; Kaplan, 1990; MacCallum, 1986; MacCallum et al., 1992; Yoon & Kim, 2014). Sampling error alone can cause such exploratory modifications to yield models that do not generalize well (MacCallum, 1986; MacCallum et al., 1992). In practice, models are typically approximations rather than perfect representatives of true data-generating processes (i.e., *approximation error*; MacCallum, 2003), in which case score test statistics can be expected to have some bias, which can then exacerbate the effect of sampling error on "capitalizing on chance" (MacCallum et al., 1992).

[8] Score tests in both `lavaan` (for complete data or FIML) and `semTools` (for multiple imputations) are calculated using expected information by default, regardless of the information matrix used to obtain *SE* estimates.

The two-stage nature of multiple imputation makes it different from FIML when it comes to the impact of approximation and sampling error. As noted previously, the first stage of multiple imputation estimates a saturated model and uses the estimates from this model to define distributions of missing values. The unstructured nature of the imputation model suggests that it may not be subject to approximation error, but our simulation results suggest that the additional layer of estimation increases noise when the sample size is small (e.g., the imputation score test was underpowered at $N = 50$). From this, there is probably no reason to expect that multiple imputation would do better than FIML in a specification search, but it is important to study whether it would do worse. An anonymous reviewer insightfully suggested that the fraction of missing information (FMI; see Graham, Olchowski, & Gilreath, 2007; Schafer, 1997) might be a useful diagnostic to consider in the context of a model modification exercise. The FMI is an intuitive quantity that captures the proportional increase in the sampling variation of an estimate or test statistic due to missing data (the FMI is largely a function of the missing data rate, but it also depends on the correlations among the variables and the missing data mechanism, among other things). The idea is that score tests or EPCs with high FMI values are more likely to capitalize on chance because the missing values substantially increase error.

We outline an FMI for the score test in Section B of the online supplement (Equation B14), which is implemented in `semTools` and demonstrated in Section C of the online supplement (subsection "Obtaining Missing-Data Diagnostics"). In the real-data example above, missing data accounted for FMI = 10% additional uncertainty in the pooled information matrix, and by extension the 3-*df* score test calculated from it. The univariate tests showed that the tested parameter with the highest modification index (199.40) and EPC (0.28) involved the variables with > 16% missing data (i.e., the covariance between the ninth and tenth indicators' residuals).

This test had a substantial FMI (25.5%), whereas the other two tested parameters with the highest modification indices (i.e., the covariances between the second and fourth indicators' and between the second and ninth indicators' residuals) had FMI < 3%. If these constituted exploratory analyses, the substantial FMI associated with the largest modification index should serve to warn the researcher that the need to validate the freed parameter in a new data set would be even greater than if the same decision were made using complete data. Savalei and Rhemtulla (2012) show how to compute FMI for parameter estimates from a FIML analysis, so similarly extending this score-test diagnostic to FIML might help adjudicate model modification decisions in both frameworks.

The intersection of model misspecification and missing data is certainly an area that could benefit from methodological research. Modification indices have constituted only one common method of exploratory model modification; other more recently proposed methods include exploratory SEM (ESEM; Asparouhouv & Muthén, 2009), SEM trees (Brandmaier, von Oertzen, McArdle, & Lindenberger, 2013), and regularized SEM (RegSEM; Jacobucci, Grimm, & McArdle, 2016). ESEM involves specifying a hypothesized restricted model with multiple-indicator constructs, and employing an estimation algorithm that allows cross-loadings to be freed in an exploratory, data-driven manner in order to improve the fit of the model. SEM trees also begin with a restricted, single-group SEM that is then recursively fitted to multiple groups defined in an exploratory fashion by selecting from a pool of candidate covariates used to "split" the sample. RegSEM[9] employs penalization (e.g., ridge or lasso) to select variables and effects to include in an exploratory model-building process. Whereas FIML could ease the application of any of these procedures to incomplete data, multiple imputation could exacerbate any existing

[9] The implementation in the R package `regsem` does not currently accommodate missing data, but it is possible in principle.

computational insensitivity (e.g., cross-validation and bootstrapping with SEM trees) and potentially complicate their application, especially if different parameters or covariates were chosen in different imputations. In the latter case, a useful generalizability diagnostic could be to only select parameters or covariates that are consistently chosen across imputations similar to how bootstrapping and random forests improve the generalizability of classification trees.

The simulation study presented here has a number of limitations. Most importantly, we only considered the "ideal" case of multivariate normal data generated under a MAR mechanism, and we restricted our attention to one latent variable model and a subset of models nested within this model (Figure 2). This was done in order to assess the properties of the multiple imputation score test under ideal conditions and compare it to established alternatives. Thus, these results may not directly generalize to non-normal data, different missing data mechanisms, or other classes of models. For instance, future research comparing the relative performance of the observed and expected information matrices under less ideal conditions would be valuable, as prior research has shown that observed information may be superior when normality is violated (Efron & Hinkley, 1978; Maydeu-Olivares, 2017). Multiple imputation is used in a variety of statistical contexts, and although we expect the current results to generalize well to linear SEMs with normally distributed MAR data, future work is needed to determine the statistical properties of multiple imputation tests, including the score test, in other contexts. Within psychology, these tests can be extended to categorical data analysis, including categorical factor analysis and structural equation modeling, as well as random coefficient modeling (e.g., growth curve models, multilevel models).

In summary, we have introduced a score test for multiply imputed data for use in model modification in structural equation modeling. We explained the statistical underpinnings of the

test, which serves as a useful guide for methodological trainees and applied researchers, and we demonstrated the superior performance of the test to the currently available FIML score test using a large simulation study. Lastly, we demonstrated the application of the multiple imputation score test in practice; in conjunction with this demonstration, we have provided detailed `R` scripts using the package `semTools` which researchers can adapt to implement the multiple imputation score test in their own data. This will be a valuable tool for structural equation modeling, the behavioral sciences in general, and other fields of science (econometrics, biomedicine) in which multiple imputation is routinely used to account for missing data.

References

Allison, P. D. (1987). Estimation of linear models with incomplete data. *Sociological Methodology, 17*, 71-103.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Associates.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16*(3), 397-438. doi:10.1080/10705510903008204

Asparouhov, T., & Muthén, B. (2010). Multiple imputation with Mplus. Retrieved from www.statmodel.com/download/Imputations7.pdf website:

Barnard, J., & Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika, 86*(4), 948-955. doi:DOI 10.1093/biomet/86.4.948

Bentler, P. M., & Bonett, D. G. (1980). Significance Tests and Goodness of Fit in the Analysis of Covariance-Structures. *Psychological Bulletin, 88*(3), 588-606. doi:Doi 10.1037/0033-2909.107.2.238

Bollen, K. A. (1989). *Structural equations with latent variables.* New York, NY: Wiley.

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods, 18*(1), 71–86. doi:10.1037/a0030001

Brown, T. A. (2014). *Confirmatory factor analysis for applied research*: Guilford Publications.

Buse, A. (1982). The Likelihood Ratio, Wald, and Lagrange Multiplier Tests - an Expository Note. *American Statistician, 36*(3), 153-157. doi:Doi 10.2307/2683166

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456.

Byron, R. P. (1972). Testing for misspecification in econometric systems using full information. *International Economic Review, 13*(3), 745–756. doi:10.2307/2525854

Carpenter, J. R., & Kenward, M. G. (2013). *Multiple imputation and its application*. West Sussex, UK: Wiley.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (Vol. 2). Pacific Grove, CA: Duxbury.

Chou, C. P., & Bentler, P. M. (1990). Model Modification in Covariance Structure Modeling - a Comparison among Likelihood Ratio, Lagrange Multiplier, and Wald Tests. *Multivariate Behavioral Research, 25*(1), 115-136. doi:DOI 10.1207/s15327906mbr2501_13

Chou, C. P., & Huh, J. (2012). Model modification in structural equation modeling. In R. Hoyle (Ed.), *Handbook of Structural Equation Modeling*, (pp. 232–246). New York, NY: Guilford.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*(4), 330-351. doi:dx.doi.org/10.1037/1082-989X.6.4.330

Efron, B., & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika, 65*(3), 457-483.

Eliason, S. R. (1993). *Maximum likelihood estimation: Logic and practice*. Newbury Park, CA: Sage.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Enders, C. K., Du, H., & Keller, B. T. (in press). A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and other nonlinear terms. *Psychological Methods*.

Enders, C. K., Keller, B. T., & Levy, R. (2018). A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychological Methods, 23*(2), 298-317. doi:10.1037/met0000148

Enders, C. K., & Mansolf, M. (2018). Assessing the fit of structural equation models with multiply imputed data. *Psychological Methods, 23*(1), 76-93. doi:10.1037/met0000102

Erler, N. S., Rizopoulos, D., Jaddoe, V. W., Franco, O. H., & Lesaffre, E. M. (2019). Bayesian imputation of time-varying covariates in linear mixed models. *Statistical Methods in Medical Research, 28*, 555-568. doi:10.1177/0962280217730851

Erler, N. S., Rizopoulos, D., Rosmalen, J., Jaddoe, V. W., Franco, O. H., & Lesaffre, E. M. (2016). Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian approach. *Statistics in Medicine, 35*(17), 2955-2974. doi:10.1002/sim.6944

Ferron, J. M., & Hess, M. R. (2007). Estimation in SEM: A concrete example. *Journal of Educational and Behavioral Statistics, 32*(1), 110-120. doi:10.3102/1076998606298025

Freedman, D. A. (2007). How can the score test be inconsistent? *The American Statistician, 61*(4), 291-295.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457-472. doi:10.1214/ss/1177011136

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2018). Package 'mvtnorm'. Retrieved from http://CRAN.R-project.org/package=mvtnorm website:

Godfrey, L. G. (1996). Misspecification tests and their uses in econometrics. *Journal of Statistical Planning and Inference*, *49*(2), 241-260.

Graham, J. W. (2003). Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling, 10*(1), 80-100. doi:Doi 10.1207/S15328007sem1001_4

Graham, J. W. (2012). *Missing data: Analysis and design*. New York: Springer.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prev Sci, 8*(3), 206-213. doi:10.1007/s11121-007-0070-9

Greene, W. H. (2012). Econometric analysis (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Grund, S., Ludtke, O., & Robitzsch, A. (2016). Pooling ANOVA Results From Multiply Imputed Datasets A Simulation Study. *Methodology-European Journal of Research Methods for the Behavioral and Social Sciences, 12*(3), 75-88. doi:10.1027/1614-2241/a000111

Honaker, J., King, G., & Blackwell, M. (2018). Package 'Amelia'. Retrieved from cran.r-project.org/web/packages/Amelia/ website:

Huang, C., & Dong, N. (2012). Factor structures of the Rosenberg self-esteem scale. *European Journal of Psychological Assessment*.

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling, 23*(4), 555-566. doi:10.1080/10705511.2016.1154793

Jaffrézic, F., White, I. M., & Thompson, R. (2003). Use of the score test as a goodness-of-fit measure of the covariance structure in genetic analysis of longitudinal data. *Genetics Selection Evolution*, *35*(2), 185.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2019). semTools: Useful tools for structural equation modeling (version 0.5-1.918). https://CRAN.R-project.org/package=semTools.

Kaplan, D. (1989). Model modification in covariance structure analysis: Application of the expected parameter change statistic. *Multivariate Behavioral Research, 24*(3), 285-305.

Kaplan, D. (1990). Evaluating and modifying covariance structure models: A review and recommendation. *Multivariate Behavioral Research, 25*(2), 137-155.

Keller, B. T., & Enders, C. K. (2019). Blimp User's Guide (Version 2). Retrieved from

Kenward, M. G., & Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science, 13*(3), 236-247.

Kline, R. B. (1998). *Principles and practice of structural equation modeling.* New York: Guilford.

Kwok, O.-M., Luo, W., & West, S. G. (2010). Using modification indexes to detect turning points in longitudinal data: A Monte Carlo study. *Structural Equation Modeling, 17*(2), 216-240.

Lee, T., & Cai, L. (2012). Alternative Multiple Imputation Inference for Mean and Covariance Structure Modeling. *Journal of Educational and Behavioral Statistics, 37*(6), 675-702. doi:10.3102/1076998612458320

Li, K. H., Raghunathan, T. E., & Rubin, D. B. (1991). Large-Sample Significance Levels from Multiply Imputed Data Using Moment-Based Statistics and an F-Reference Distribution.

Journal of the American Statistical Association, 86(416), 1065-1073. doi:Doi 10.2307/2290525

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Hoboken, NJ: Wiley.

Liu, J. C., Gelman, A., Hill, J., Su, Y. S., & Kropko, J. (2014). On the stationary distribution of iterative imputations. *Biometrika, 101*(1), 155-173. doi:10.1093/biomet/ast044

Liu, Y., & Enders, C. K. (2016). *Evaluation of multi-parameter test statistics for multiple imputation.*

MacCallum, R. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin, 100*(1), 107.

MacCallum, R. C. (2003). 2001 presidential address: Working with imperfect models. *Multivariate Behavioral Research, 38*(1), 113-139.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychological Bulletin, 111*(3), 490.

Maydeu-Olivares, A. (2017). Maximum likelihood estimation of structural equation models for continuous data: Standard errors and goodness of fit. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(3), 383-394.

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *The Journal of Mathematical Sociology, 4*(1), 103-120. doi:10.1080/0022250x.1975.9989847

Meng, X. L., & Rubin, D. B. (1992). Performing Likelihood Ratio Tests with Multiply-Imputed Data Sets. *Biometrika, 79*(1), 103-111. doi:Doi 10.2307/2337151

Morgan, B. T., Palmer, K. J., & Ridout, M. S. (2007). Negative score test statistic. *The American Statistician, 61*(4), 285-288.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52*(3), 431-462.

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide. Eighth edition.* Los Angeles, CA: Muthén & Muthén.

Nevitt, J., & Hancock, G. R. (2004). Evaluating small sample approaches for model test statistics in structural equation modeling. *Multivariate Behavioral Research, 39*(3), 439-478. doi:10.1207/S15327906MBR3903_3

Oberski, D. L. (2014). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis, 22*(1), 45-60.

Quartagno, M., & Carpenter, J. (2018). Package 'jomo'. Retrieved from cran.r-project.org/web/packages/jomo/ website:

Rao, C. R. (1948). *Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation.* Paper presented at the Mathematical Proceedings of the Cambridge Philosophical Society.

Reise, S. P., Kim, D. S., Mansolf, M., & Widaman, K. F. (2016). Is the bifactor model a better model or is it just better at modeling implausible responses? Application of iteratively reweighted least squares to the Rosenberg Self-Esteem Scale. *Multivariate Behavioral Research, 51*(6), 818-838.

Reiter, J. P. (2007). Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. *Biometrika, 94*(2), 502-508. doi:10.1093/biomet/asm028

Reiter, J. P., & Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association, 102*(480), 1462-1471. doi:10.1198/016214507000000932

Rosenberg, M. (1965). Rosenberg self-esteem scale (RSE). *Acceptance and commitment therapy. Measures package, 61*, 52.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1-36.

Rubin, D. B. (1987). *Multiple imputation for nonrespose in surveys*. New York: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 91*(434), 473-489.

Saris, W. E., Satorra, A., & Sörbom, D. (1987). The detection and correction of specification errors in structural equation models. *Sociological Methodology*, 105-129.

Saris, W. E., Satorra, A., & Van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling, 16*(4), 561-582.

Sato, S., Ueki, M., & Alzheimer's Disease Neuroimaging Initiative. (2018). Fast score test with global null estimation regardless of missing genotypes. *PloS one*, *13*(7), e0199692.

Savalei, V. (2010a). Expected versus observed information in SEM with incomplete normal and nonnormal data. *Psychological Methods, 15*(4), 352-367. doi:10.1037/a0020143

Savalei, V. (2010b). Small Sample Statistics for Incomplete Nonnormal Data: Extensions of Complete Data Formulae and a Monte Carlo Comparison. *Structural Equation Modeling-a Multidisciplinary Journal, 17*(2), 241-264. doi:10.1080/10705511003659375

Savalei, V., & Bentler, P. M. (2009). A Two-Stage Approach to Missing Data: Theory and Application to Auxiliary Variables. *Structural Equation Modeling-a Multidisciplinary Journal, 16*(3), 477-497. doi:10.1080/10705510903008238

Savalei, V., & Falk, C. F. (2014). Robust Two-Stage Approach Outperforms Robust Full Information Maximum Likelihood With Incomplete Nonnormal Data. *Structural Equation Modeling-a Multidisciplinary Journal, 21*(2), 280-302. doi:10.1080/10705511.2014.882692

Savalei, V., & Rhemtulla, M. (2012). On obtaining estimates of the fraction of missing information from full information maximum likelihood. *Structural Equation Modeling: A Multidisciplinary Journal, 19*(3), 477-494.

Savalei, V., & Rhemtulla, M. (2014). *Two-stage estimator for models with composites or parcels when data are missing at the item level*. Paper presented at the Society of Multivariate Experimental Psychology, Nashville, TN.

Savalei, V., & Yuan, K. H. (2009). On the Model-Based Bootstrap With Missing Data: Obtaining a P-Value for a Test of Exact Fit. *Multivariate Behav Res, 44*(6), 741-763. doi:10.1080/00273170903333590

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York: Chapman & Hall.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177. doi:10.1037//1082-989x.7.2.147

Schafer, J. L., & Olsen, M. K. (1998). Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. *Multivariate Behav Res, 33*(4), 545-571. doi:10.1207/s15327906mbr3304_5

Silvey, S. (1975). *Statistical Inference* (Vol. 7): CRC Press.

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods, 6*(4), 317-329. doi:10.1037//1082-989x.6.4.317

Sörbom, D. (1989). Model modification. *Psychometrika, 54*(3), 371-384.

Spanos, A. (1999). *Probability theory and statistical inference: econometric modeling with observational data*: Cambridge University Press.

Steele, R. J., Wang, N., & Raftery, A. E. (2010). Inference from multiple imputation for missing data using mixtures of normals. *Statistical Methodology, 7*, 351-365.

van Buuren, S. (2012). *Flexible imputation of missing data*. New York: Chapman & Hall.

Van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation, 76*(12), 1049-1064. doi:10.1080/10629360600810434

van Buuren, S., Groothuis-Oudshoorn, K., Robitzsch, A., Vink, G., Doove, L., Jolani, S., . . . Gray, B. (2018). Package 'mice'. Retrieved from cran.r-project.org/web/packages/mice/mice.pdf website:

Verbeke, G., & Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics, 59*(2), 254-262.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society, 54*(3), 426-482.

Whittaker, T. A. (2012). Using the modification index and standardized expected parameter change for model modification. *The Journal of Experimental Education, 80*(1), 26-44.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics, 9*(1), 60-62.

Yoon, M., & Kim, E. S. (2014). A comparison of sequential and nonsequential specification searches in testing factorial invariance. *Behavior Research Methods, 46*(4), 1199-1206.

Yuan, K.-H., Tong, X., & Zhang, Z. (2014). Bias and Efficiency for SEM With Missing Data and Auxiliary Variables: Two-Stage Robust Method Versus Two-Stage ML. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(2), 178-192. doi:10.1080/10705511.2014.935750

Yuan, K. H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology 2000, Vol 30, 30*, 165-200. doi:Doi 10.1111/0081-1750.00078

Yuan, K. H., & Bentler, P. M. (2010). Consistency of Normal Distribution Based Pseudo Maximum Likelihood Estimates When Data Are Missing at Random. *Am Stat, 64*(3), 263-267. doi:10.1198/tast.2010.09203

Yuan, K. H., & Savalei, V. (2014). Consistency, bias and efficiency of the normal-distribution-based MLE: The role of auxiliary variables. *Journal of Multivariate Analysis, 124*, 353-370. doi:10.1016/j.jmva.2013.11.006

Yuan, K. H., & Zhang, Z. Y. (2012). Robust Structural Equation Modeling with Missing Data and Auxiliary Variables. *Psychometrika, 77*(4), 803-826. doi:10.1007/s11336-012-9282-4

Table 1

Structural regression coefficients for simulation study ($\beta_{14} = \beta_{24} = \beta_{34}$)

	Number of Parameters		
R^2	1	2	3
.02	.141	.088	.065
.13	.361	.224	.165
.26	.510	.316	.233

Table 2

Item Content for the Rosenberg Self-Esteem Scale

1. I feel that I am a person of worth, at least on an equal plane with others.
2. I feel that I have a number of good qualities.
3. All in all, I am inclined to feel that I am a failure. (R)
4. I am able to do things as well as most other people.
5. I feel I do not have much to be proud of. (R)
6. I take a positive attitude toward myself.
7. On the whole, I am satisfied with myself.
8. I wish I could have more respect for myself. (R)
9. I certainly feel useless at times. (R)
10. At times I think I am no good at all. (R)

Note. (R) indicates reverse worded and scored.

Table 3

Modification indices and expected parameter change statistics for the Rosenberg data.

Residual Covariance	MI	EPC
(9, 10)	199.40	0.28
(1, 2)	163.99	0.15
(6, 7)	129.69	0.15
(2, 4)	53.60	0.09
(2, 9)	38.67	-0.09

Note. MI = modification index; EPC = expected parameter change. Results are presented in decreasing order of MI, and only the five highest values of MI are presented.

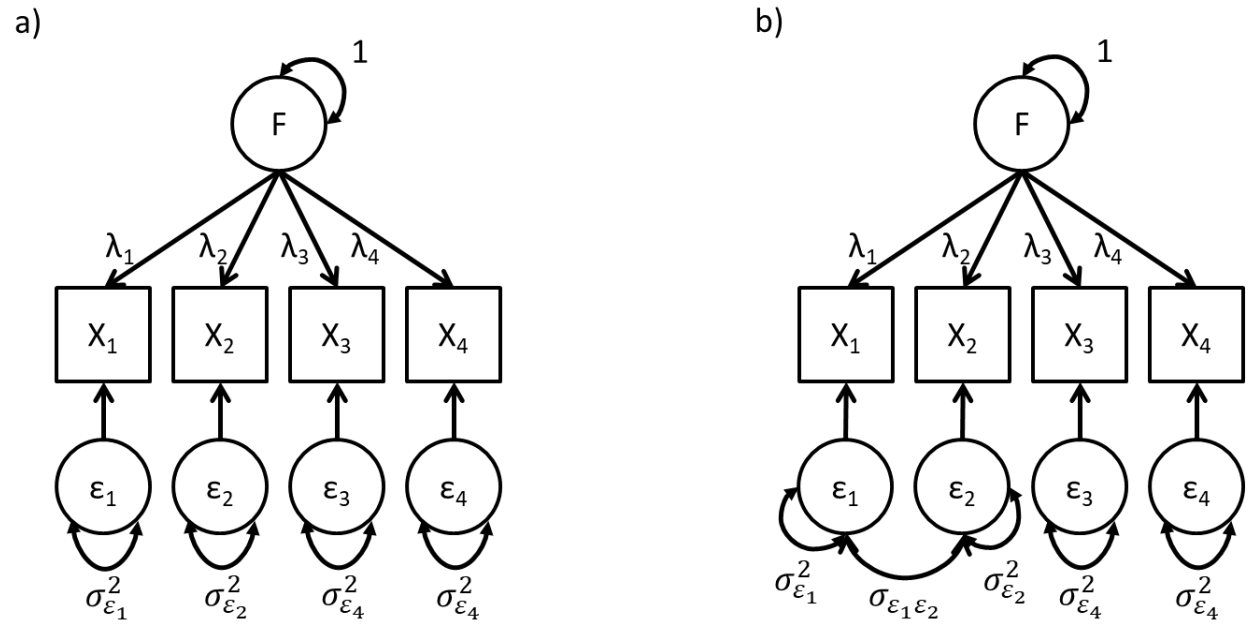

Figure 1. Path diagrams for factor analysis models. F represents the latent variable, $X_1, \ldots, X_4$ represent the four observed variables, $\lambda_1, \ldots, \lambda_4$ represent the factor loadings, and $\varepsilon_1, \ldots, \varepsilon_4$ represent the residuals for $X_1, \ldots, X_4$ after controlling for F, with $\sigma^2_{\varepsilon_2}$ $\sigma^2_{\varepsilon_1}, \ldots, \sigma^2_{\varepsilon_4}$ representing the residual variances of $\varepsilon_1, \ldots, \varepsilon_4$ and $\sigma_{\varepsilon_1\varepsilon_2}$ representing the residual covariance of ε_1 and ε_2. Figure 1a contains the restricted model, with $\sigma_{\varepsilon_1\varepsilon_2}$ constrained to zero (not estimated). Figure 1b contains the general model with $\sigma_{\varepsilon_1\varepsilon_2}$ unconstrained (freely estimated).

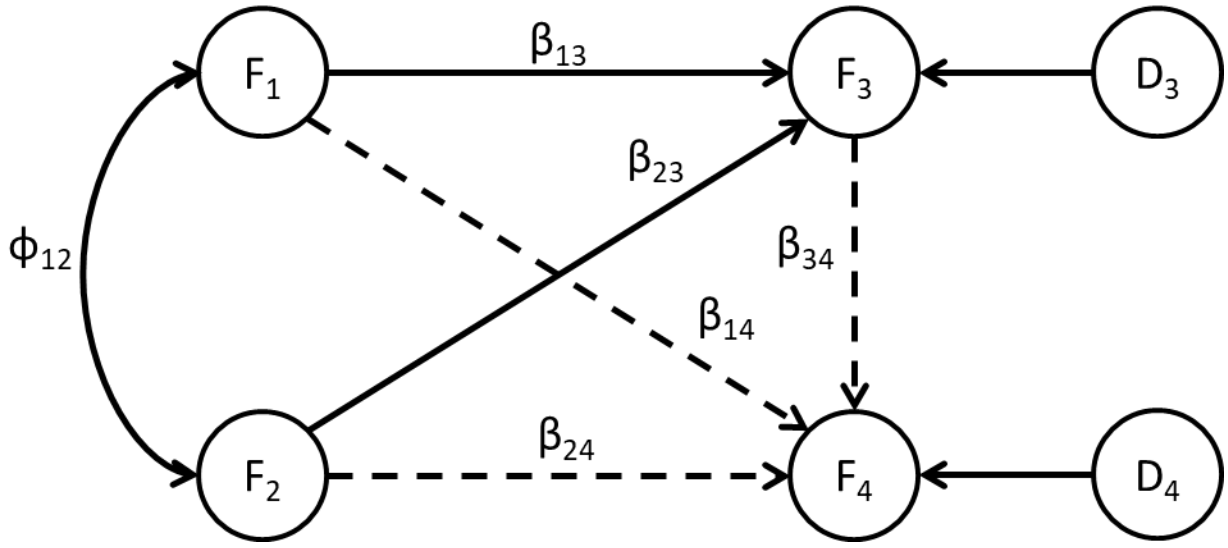

Figure 2. Latent regression model used in simulation study. Only latent variables and structural paths are shown. Dashed paths are manipulated and tested according to the simulation conditions. Adapted from Figure 3 in Chou and Bentler (1990, p. 124).

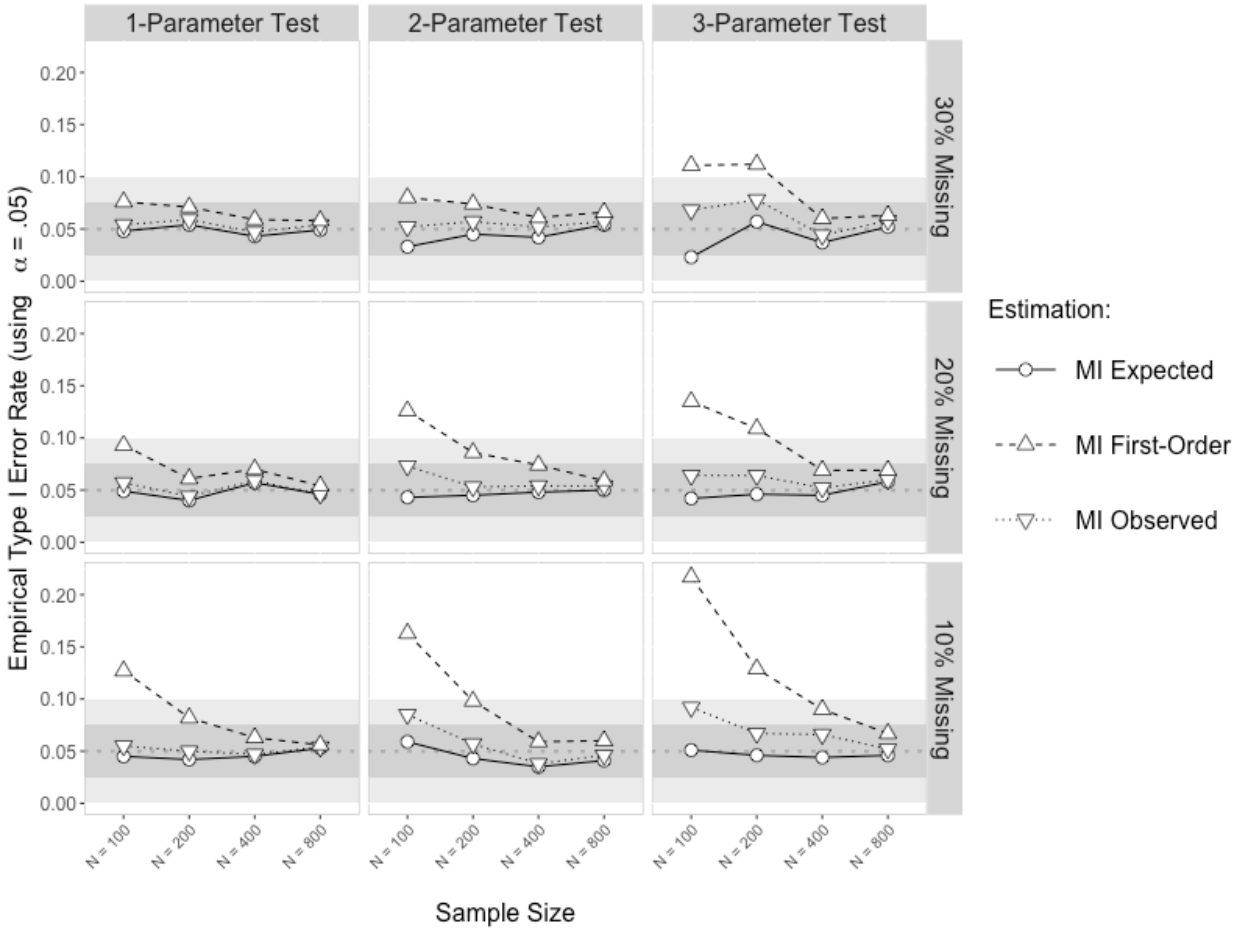

Figure 3. Empirical Type I error for multiple imputation (MI) score tests. An alpha level of .05, indicated by the dotted horizontal line, was used as the criterion for statistical significance. The dark gray shaded region indicates little to no bias in Type I error rate, the light gray shaded region indicates moderate bias in Type I error rate, and the white region indicates large bias in Type I error rate. *Expected*, *observed*, and *first-order* denote the information matrix used.

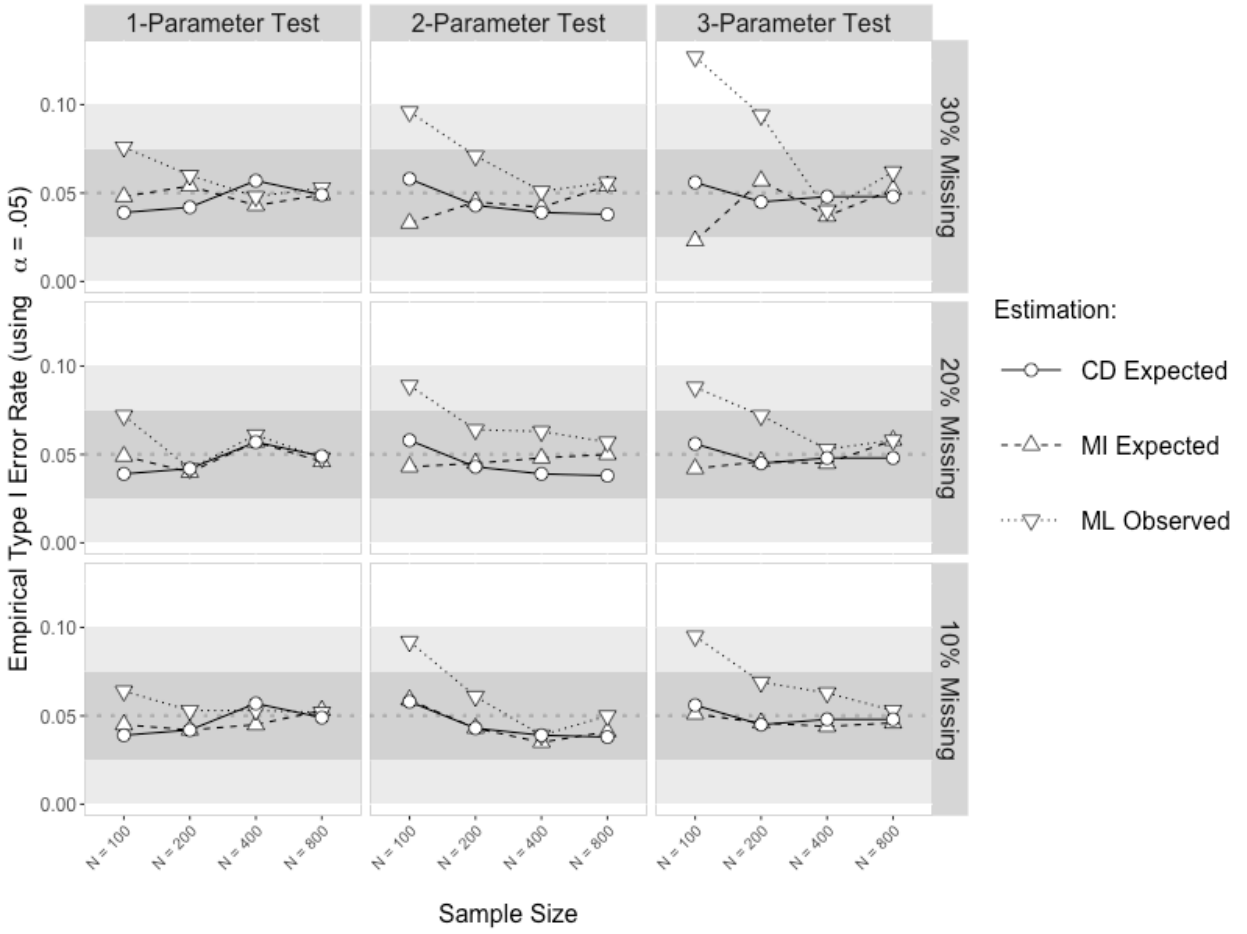

Figure 4. Empirical Type I error for multiple imputation (MI), complete-data (CD), and full-information maximum likelihood (ML) score tests. An alpha level of .05, indicated by the dotted horizontal line, was used as the criterion for statistical significance. The dark gray shaded region indicates little to no bias in Type I error rate, the light gray shaded region indicates moderate bias in Type I error rate, and the white region indicates large bias in Type I error rate. *Expected* and *observed* denote the information matrix used.

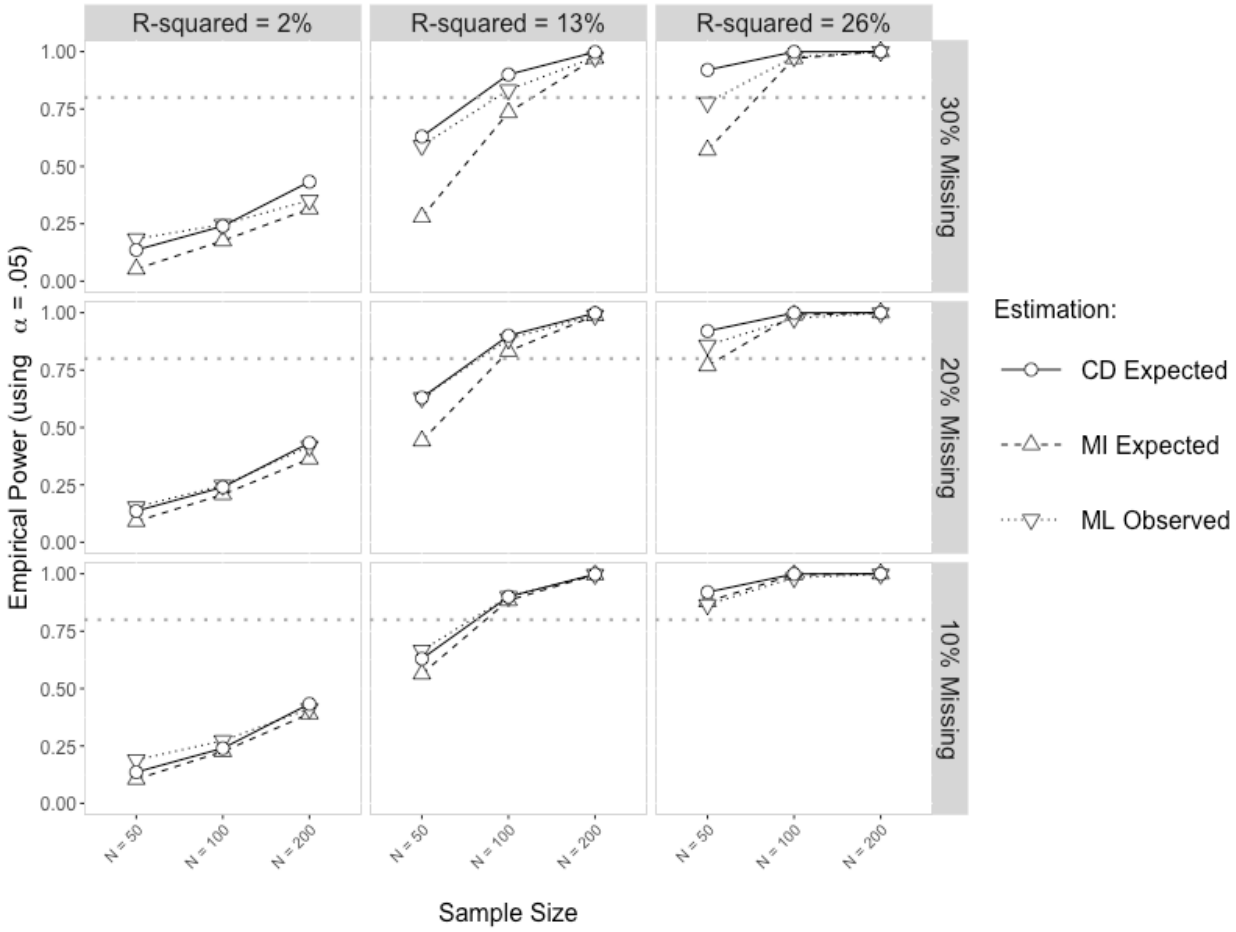

Figure 5. Empirical power for single-parameter multiple imputation (MI), complete-data (CD), and full-information maximum likelihood (ML) score tests. Power of .8 (using α = .05 as criterion for significance) is indicated by the dotted horizontal line. *Expected* and *observed* denote the information matrix used.

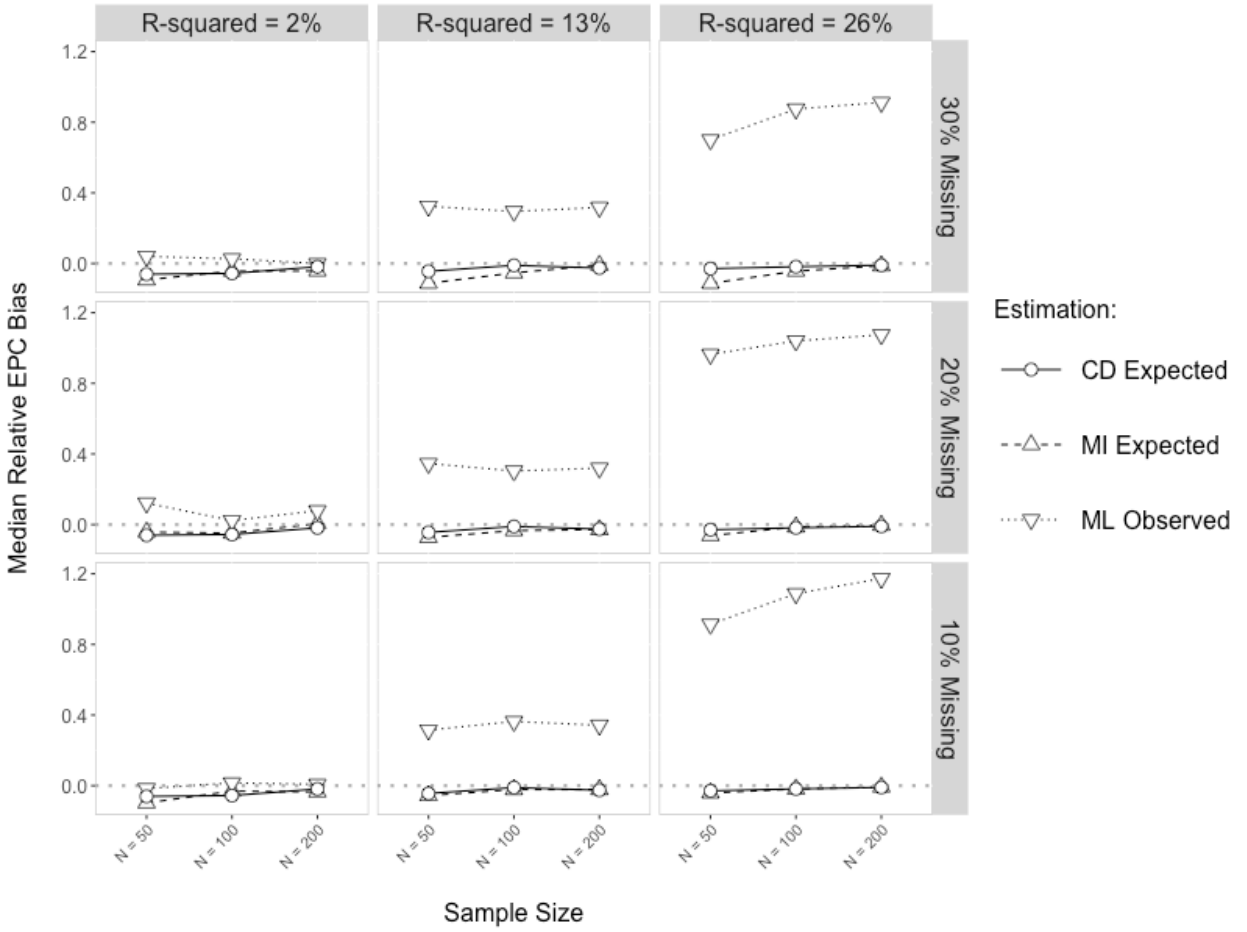

Figure 6. Median relative bias in EPC. Zero bias is indicated by the dotted horizontal line. CD = complete data; MI = multiple imputation; ML = full-information maximum likelihood. *Expected* and *observed* denote the information matrix used.

Appendix

R Syntax for the Real Data Example

```
## Rosenberg (1979) Self-Esteem Scale data provided by
##      http://personality-testing.info/_rawdata/
## R scripts that import data, impose missing values, and multiply impute the
## data can be found on this project's Open Science Framework (OSF) page:
##      https://osf.io/m2gd8/


## ------------------------
## Import the Imputed Data
## ------------------------

## stored as a single stacked data.frame with an indicator variable
allimps <- read.csv(file = "rosenimps.csv", header = FALSE,
                    col.names = c("imp", colnames(cdata)))
## extract the rows for each imputation, store as a list of imputations
impList <- lapply(1:max(allimps$imp),
                  function(m) allimps[allimps$imp == m, -1])
## NOTE: Annotated syntax on our OSF page also reverse-scores items


## ------------------------
## Analyze the Imputed Data
## ------------------------

## load semTools (at least version 0.5-1.921), which also loads lavaan
library(semTools)
## if your version is not up to date, install the development version:
##     devtools::install_github("simsem/semTools/semTools")

## specify the restricted model
(model <- paste("F =~", paste0("X", 1:10, collapse = " + ")))
## fit the model to imputed data
fit.imps <- cfa.mi(model, data = impList, std.lv = TRUE)
## obtain pooled results (estimates, fit measures)
summary(fit.imps)
fitMeasures(fit.imps)

## specify constrained parameters that could be added to the model
myResidCors <- c('X9 ~~ X10', 'X1 ~~ X2', 'X6 ~~ X7')
## request the score test, univariate tests of each parameter, and their EPCs
out.imps <- lavTestScore.mi(fit.imps, add = myResidCors, epc = TRUE,
                            test = "D1", asymptotic = TRUE)
## print multiparameter score test
out.imps$test
## print individual univariate tests, with expected parameter changes ...
out.imps$uni
## ... which are the usual modification indices (also with EPCs)
modindices.mi(fit.imps, op = "~~", minimum.value = 50, sort. = TRUE,
              test = "D1")
```